Transparent Embedded Compression in Systems-on-Chip

Bram Riemens, Principal Scientist, Research, NXP Semiconductors
bram.riemens@nxp.com
With René van der Vleuten, Pieter van der Wolf,
Geogy Jacob, Jan-Willem van de Waerdt, Johan Janssen

IEEE Workshop on Signal Processing Systems SiPS 2006
Banff, AB, Canada. October 2-4, 2006

Introduction

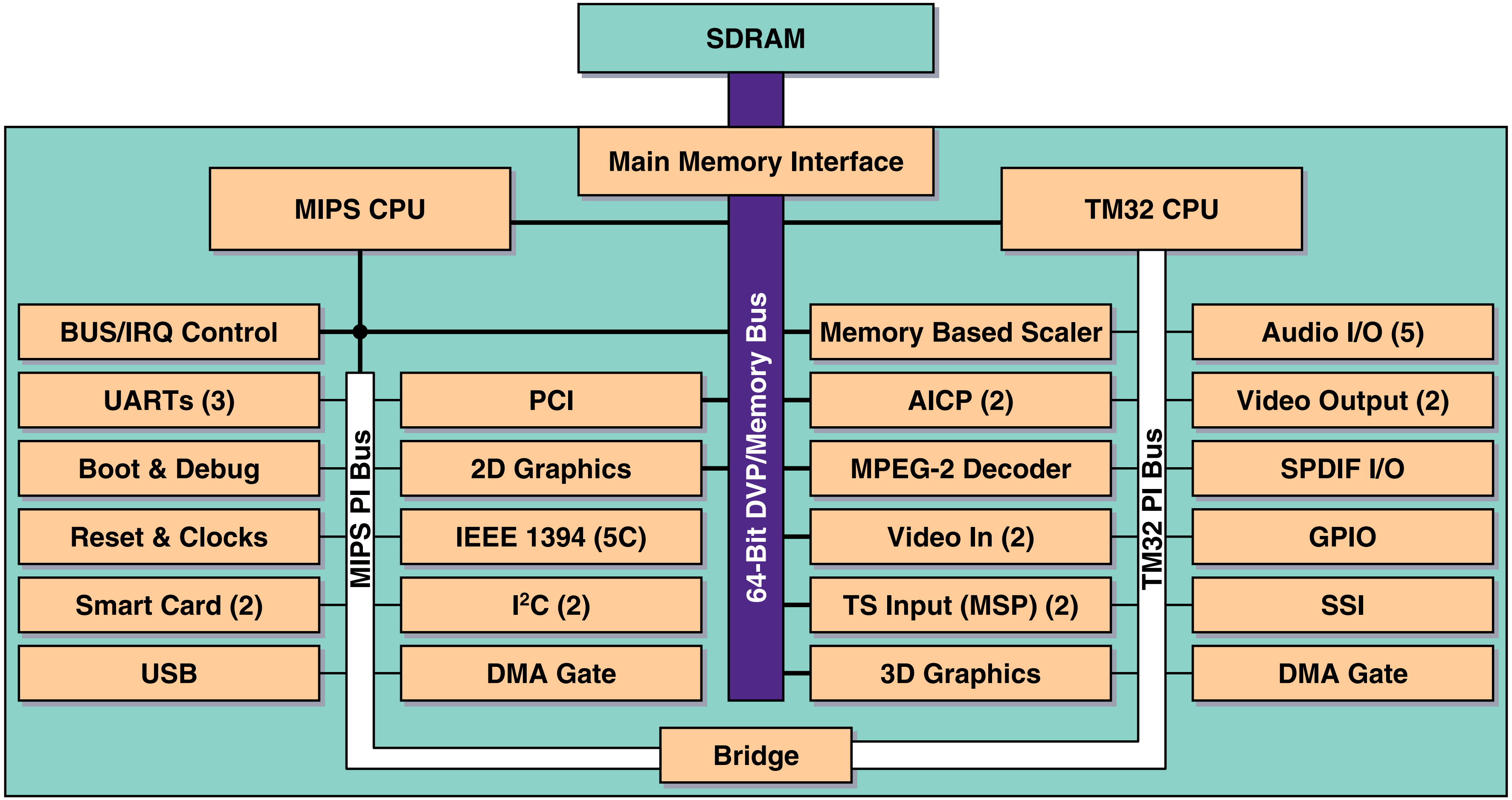
Operating principle

Architecture

Algorithm

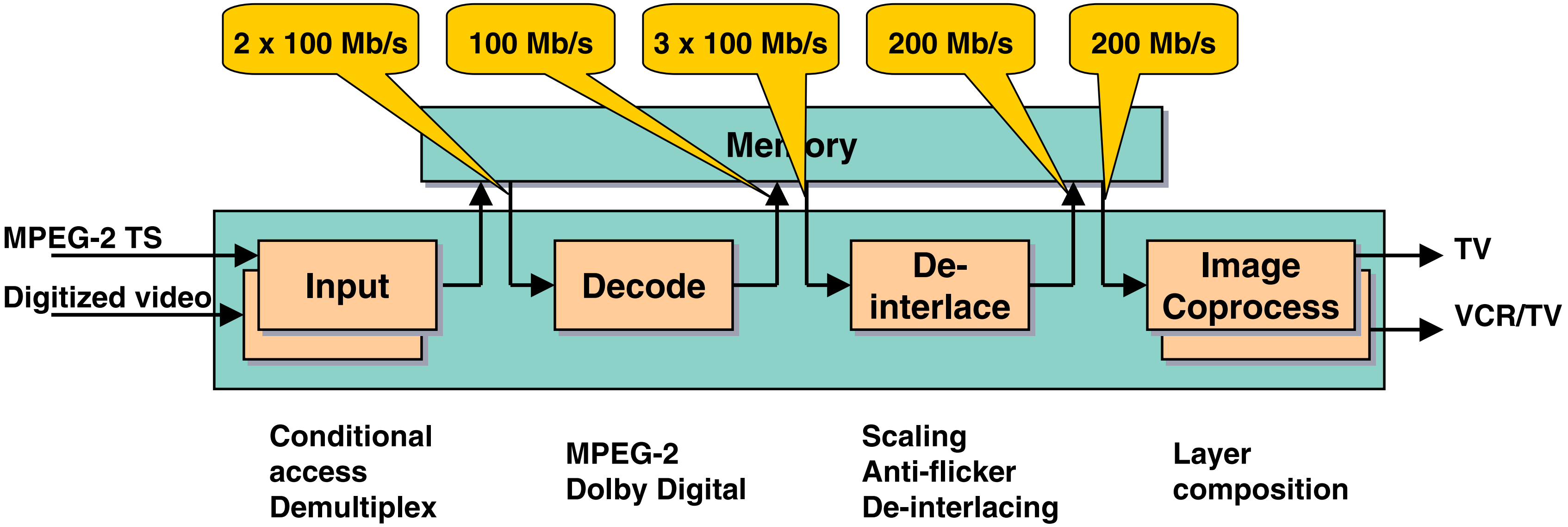
Results and Conclusions

Introduction – Existing SoC system



Introduction – Application on SoC

- Unified memory located externally from the processing chip
- Significant amount of available memory bandwidth consumed by image data
- Both hardware and software processing components
- Many applications and use-cases – variations in resource use



Introduction

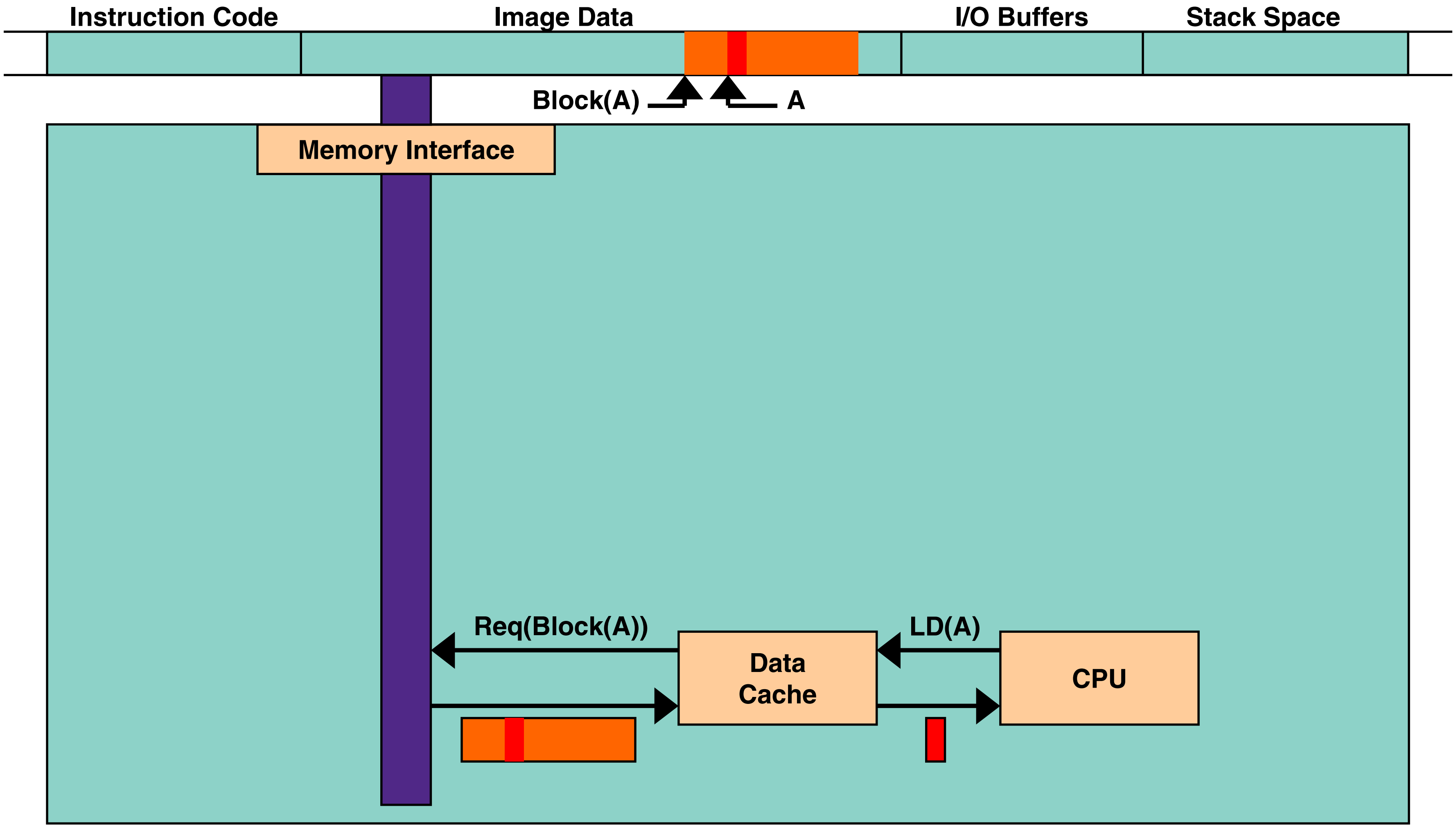
Operating principle

Architecture

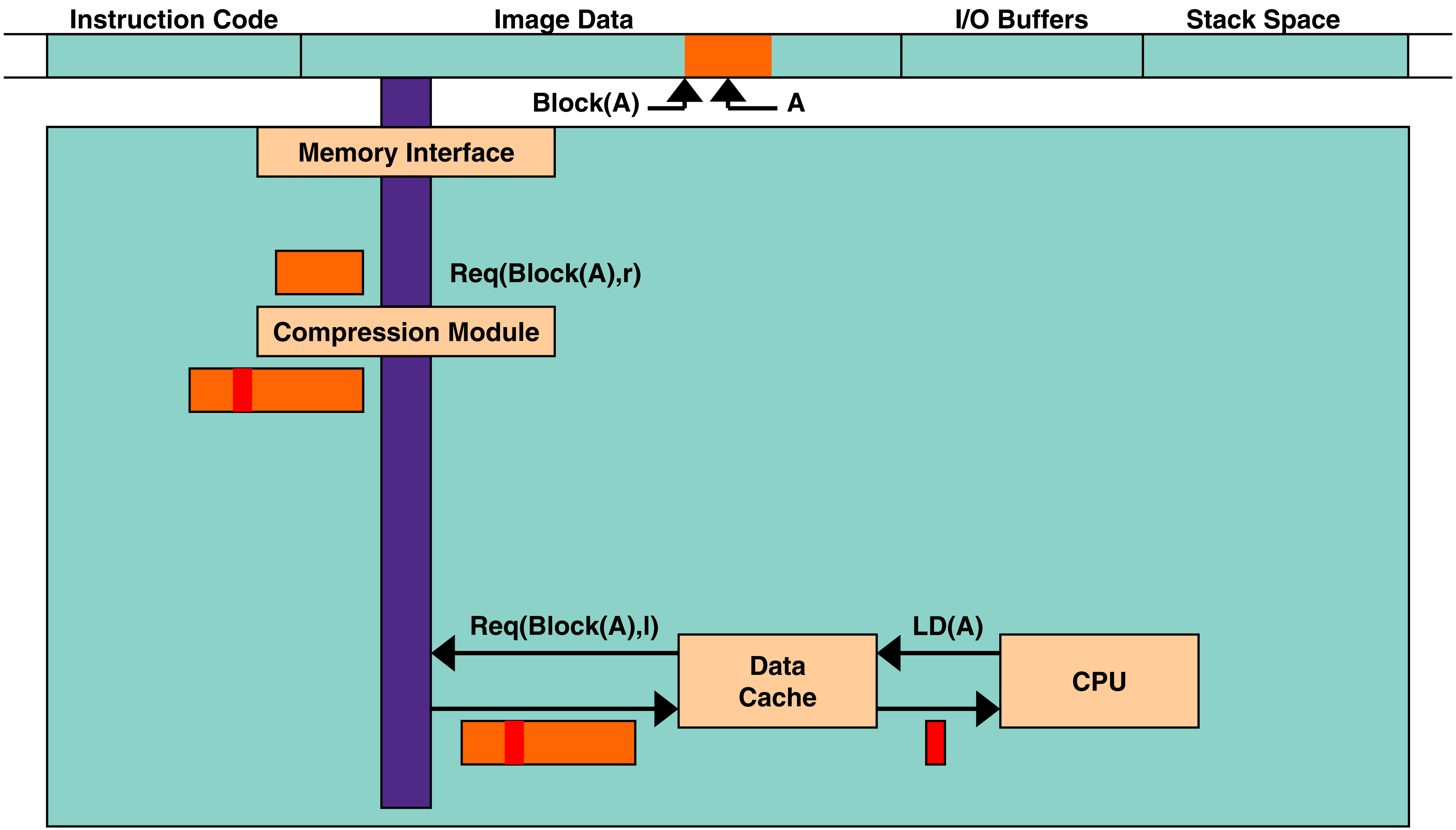
Algorithm

Results and Conclusions

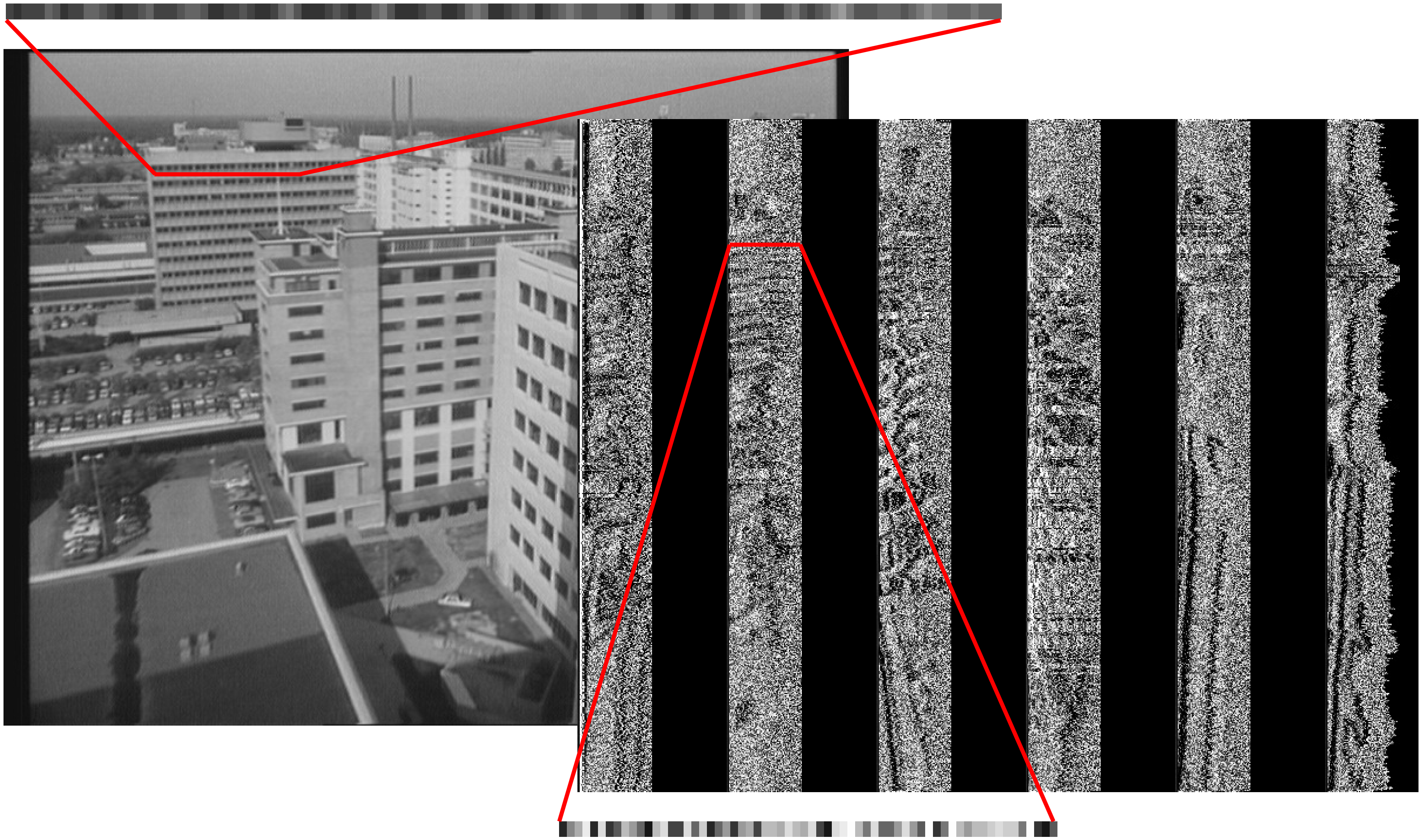
Data transfer of existing systems



Data transfer with embedded compression



Memory layout visualization



Problem definition summary

- Memory access speed improves at slower pace than processing speed
  - Bandwidth is becoming a dominant design issue
- Consumer market, high volume
  - Cost effective solutions required
- Actual bandwidth consumption is dynamic; depending on
  - Algorithms
  - Use-case (details!)
  - Image contents
  - Caching behavior
- Bandwidth limit is hard boundary
  - For real-time systems: too late is an error
- Cope with legacy hardware and software components

Solution approach

Solution: transparent embedded compression

- ***Compression***

  Reduce off-chip memory bandwidth consumed by image data

- ***Embedded***

  Compress on write, decompress on read

  So: freedom of algorithm choice (not bound to any standard)

- ***Transparent***

  No need to adapt signal processing units to add compression

  So: incorporate compression in the communication infrastructure

Introduction

Operating principle

Architecture

Algorithm

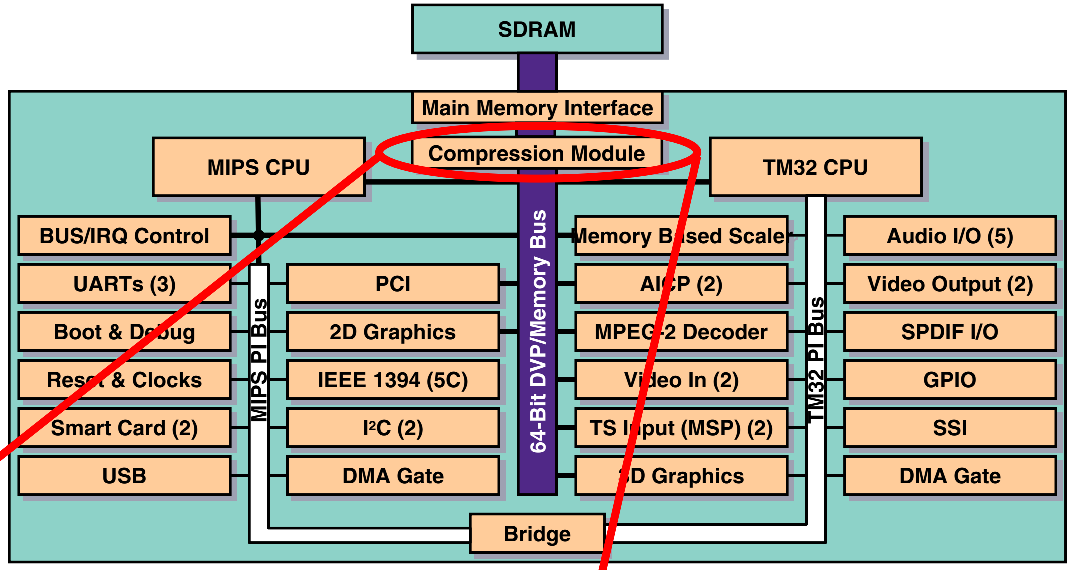
Results and Conclusions

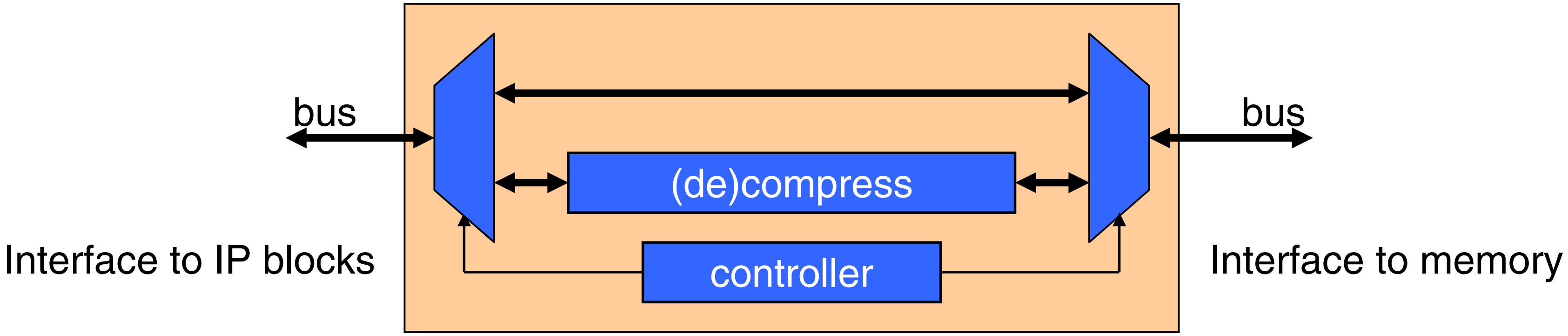
Architecture – Requirements summary

- Compression module is part of the communication infrastructure
- Operate on single data transaction to/from memory (no state)
- Latency
  - Predictable for compressed transfers
  - Minimal for other transfers

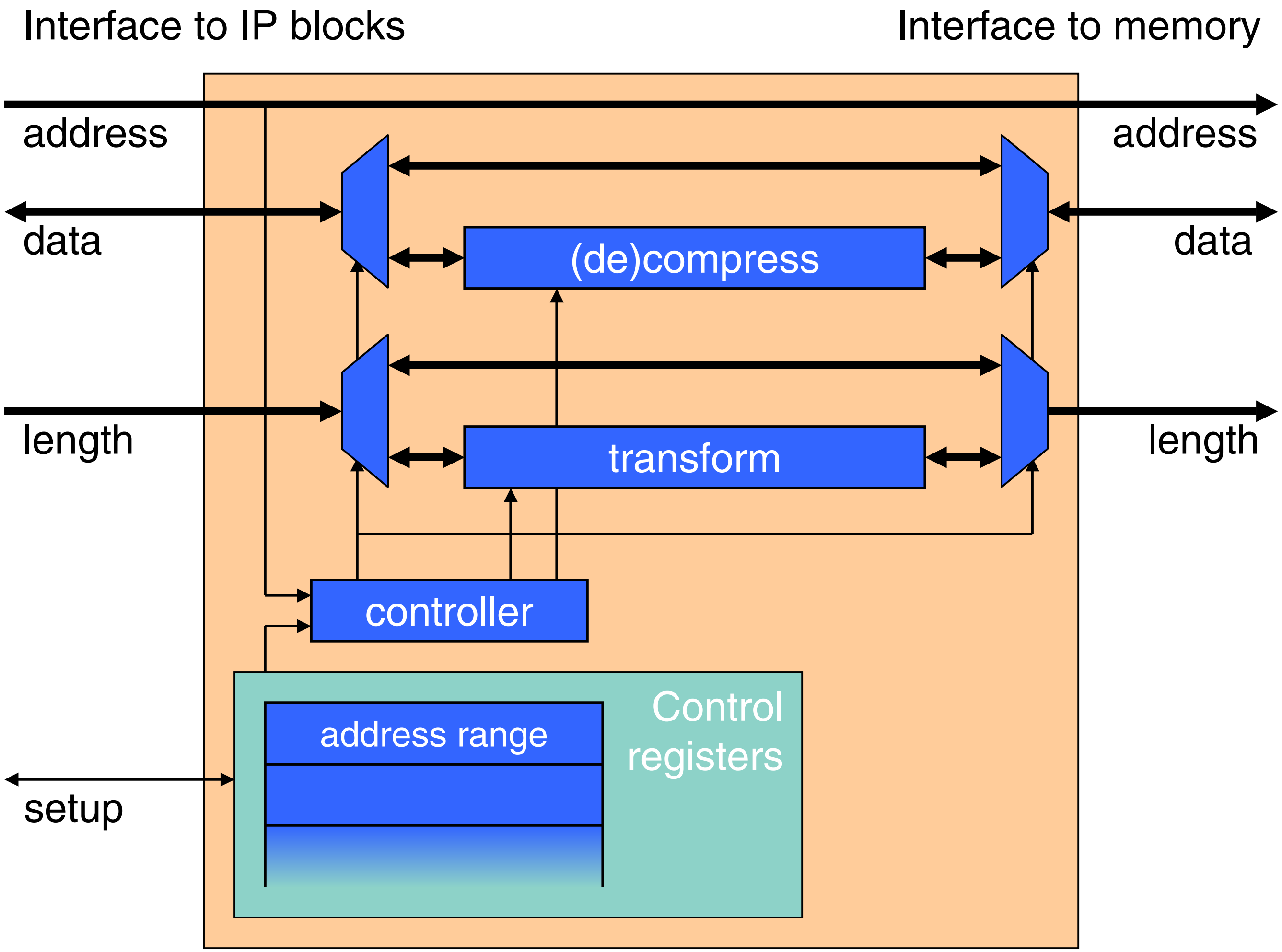
Architecture – Memory bus module

- Extra component
  in the bus hierarchy
- (De)compress image data
- Bypass other data





Architecture – Block diagram



Introduction

Operating principle

Architecture

Algorithm

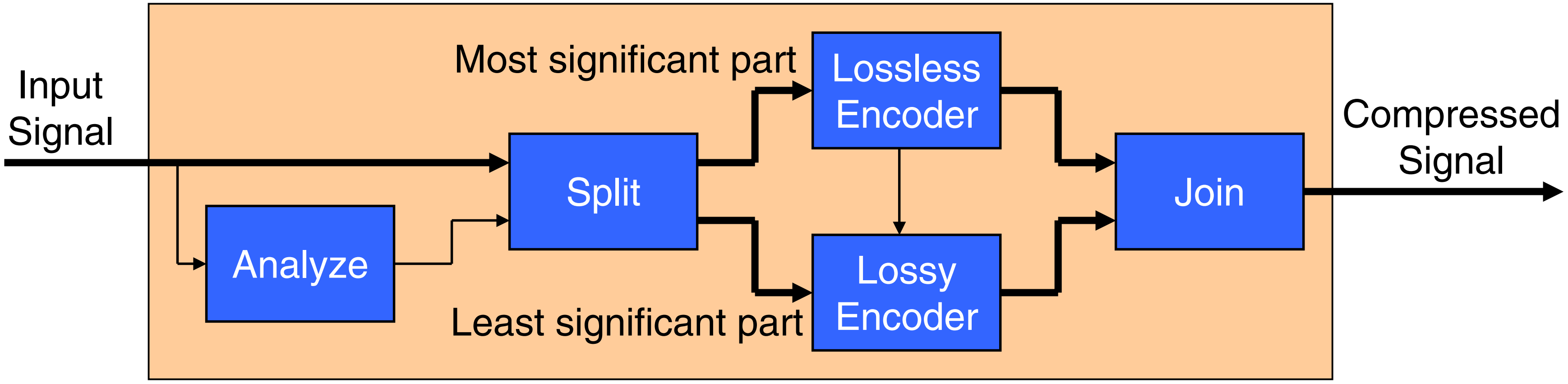
Results and Conclusions

Algorithm – Requirements summary

- Operate on single data transaction to/from memory (no state)
- Low complexity algorithm (cost)
- Emphasis on bandwidth guarantees, so lossy compression
- Compression ratio at least 1.5
- No significant impact on image quality
- Adjustable compression ratio
- Latency
  - Predictable for compressed transfers
  - Minimal for other transfers

Algorithm – Block diagram and features

- Assuming horizontally neighboring pixel values
- Dynamic split of MSB / LSB part
  - MSB: lossless, DPCM and VLC (modified “Rice” code)
  - LSB: lossy, distribute bit planes over remaining space
- Can handle:
  - Y, U/V, and R/G/B multiplexing schemes
  - Video images and graphics textures
- No further signal degradation after first compression pass
- Few control bits to pass coding information to decoder



Algorithm – Image quality verification

Extensive “torture” tests

- Large database with critical scenes
- Special test patterns, noisy data
- Repeated compression / decompression to prove viability in recursive loops
- Both 8 and 10 bits sources
- Combined with other enhancement functions (e.g. sharpness) to increase sensitivity to artifacts
- Visual inspection to obtain perceptually optimal result

For reference: PSNR on “Lena”:

– 8 bit	1.5	54.81 dB	(target use-case)
– 10bit	1.875	54.60 dB	(target use-case)
– 8 bit	2.0	45.63 dB	(fallback use-case)

Introduction

Operating principle

Architecture

Algorithm

Results and Conclusions

Results – Data

- Effective bandwidth saving
  - Compression ratio 1.5 20% - 25%
  - Compression ratio 2.0 ± 40%
- Area: ± 1 mm^2 in 90 nm CMOS
- Clock: 350 MHz
- Latency
  - 128 byte transfer: 80 cy compression; 58 cy decompression
  - 256 byte transfer: 144 cy compression; 106 cy decompression
  - Note: prefetching can hide this additional latency

Conclusions

- Transparent embedded compression in industrially relevant context
- Requirements are met at reasonable cost
  - Legacy IP (hw or sw) still applicable
  - Add-on to existing communication infrastructure
- Enable dynamic trade-off between bandwidth consumption and image quality, enabling:
  - Quality-of-service
  - De-risking of system design
  - Optimize image quality over processing chain with minimal bandwidth use
  - System differentiation without SoC redesign (apply different RAM speed ratings)